DOCUMENT RESUME

ED 129 871 TM 005 672

AUTHOR Merz, William R.
TITLE Estimating Bias in Test Items Utilizing Principal Components Analysis and the General Linear Solution.
PUB DATE Apr 76
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 92-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS American Indians; *Analysis of Variance; Anglo Americans; Elementary School Students; Ethnic Groups; *Factor Analysis; Factor Structure; *Item Analysis; Mexican Americans; Multiple Regression Analysis; Negroes; Orthogonal Rotation; Projective Tests; Racial Discrimination; Statistical Analysis; *Test Bias
IDENTIFIERS Goodenough Harris Drawing Test; Principal Components Analysis; *Test Items

ABSTRACT

A number of methods have been used to identify potentially biased items within a test. These methods examine one item at a time and do not deal with the complex interrelationships among items or among items and the potentially biasing elements. The use of multivariate procedures to assess whether or not items are biased and to obtain clues about the source of the bias are demonstrated here. A total of 1,294 six and seven year old children from five ethnic groups took the Goodenough-Harris Drawing Test. Principal components analysis and analysis of variance were performed on the results. Other analysis methods are suggested, and are presently being studied. (Author/BW)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED129871

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ESTIMATING BIAS IN TEST ITEMS

UTILIZING PRINCIPAL COMPONENTS ANALYSIS AND THE GENERAL LINEAR SOLUTION

William R. Merz, Ph.D.
Associate Professor of Behavioral Sciences in Education
California State University at Sacramento

The issue of bias in testing has become a major concern for professionals of many disciplines. Educators, personnel workers, psychologists, psychometricians, and sociologists not only have studied the issues of bias in testing and the fair use of tests but also have been directed by the courts and by governmental regulatory agencies to construct tests free from bias before using them as evaluation and selection devices.

Recent efforts in examining tests for bias have focused on the individual test items themselves rather than on the total scores obtained by individuals on the test (Echternacht, 1974, Gree.., 1975, 1973, 1971; Green and Draper, 1972; Merz, 1974, 1973; Ozenne, Van Gelder, and Cohen, 1974; Scheuneman, 1975). These studies have examined the bias of test items in the absence of external criteria; that is, they have defined bias as item by group interaction.

The investigators cited above approached the estimation of item bias in different ways. Echternacht (1974) began with item p-values for each independent group of interest transforming the item p-values to delta values. The delta differences were obtained for each pair of groups. These differences were plotted on normal probability paper along with the line representing a hypothetical normal distribution with the obtained mean and variance of the differences as parameters. Confidence bands were then drawn around the line. Biased items were defined as those falling outside of the bands.

TM005 672

Ozenne, Van Gelder and Cohen (1974) suggested a two-step procedure. First, item difficulty levels from most difficult to most easy were plotted using one of the identified groups as a reference. Visual examination of the shape of the plots indicated item by group interaction when the uniformity of the curves was disturbed. Second, point biserial correlations for item and total score were computed for each independent group. These correlations were compared to identify which items did not contribute to total score for specific sub-groups. Items were identified as potentially biased by expert judgement based on the results of the two methods of analysis.

CTB/McGraw-Hill (1974) utilized point biserial correlations to assess the bias of items in the Comprehensive Tests of Basic Skills, Form S. An absolute difference between pairs of .2 was used to identify items which were biased. Scheuneman (1975) employed Chi-Square to identify potentially biased items in trial versions of the Metropolitan Readiness Tests. The frequency of individuals in each group of interest who got an item correct was tabulated into one of four categories; these categories might represent those whose total score fell into each of the four quartiles. Then, proportions of those who scored successfully on each item were compared across quartiles with Chi-Square.

Each of the methods described examines one item at a time. None deal with the complex interrelationship among items or among items and the potentially biasing elements. Each of the methods identifies an item as biased without giving clues regarding the source of that bias. Determining the source of the bias is left to expert analysis of the task and the content included in the item. Groups are compared a pair at a time. By examining one variable at a time, interactions between and among characteristics cannot be assessed. It is the contention of this writer that methods for assessing item bias must take

into account the complex interrelationships among items within a test and must provide clues about the source of that bias. In order to do this, multivariate techniques are necessary.

Two multivariate approaches are proposed in this paper. Both take into account the complex relationships among items within a test and among the items and the potential sources of bias. Both provide clues about the source of the bias. Kerlinger and Pedhazur (1973) claimed that one method is the equivalent of the other. The first method employes principal components analysis to reduce the item intercorrelation matrix and analysis of variance to examine resultant factor scores. The second method employs principal components, too, to reduce the item intercorrelation matrix and multiple regression analysis with subjects' group memberships as dummy predictor variables and factor scores as criterion variables. The steps are outlined below.

Steps
Common to Both
Methods

1. Compute the item intercorrelation matrix.
2. Reduce that matrix with principal components analysis. Use the Scree Principle to determine how many factors to extract. Usually setting the eigenvalue criterion at 1.0 will result in extraction of no more factors than 1/3 the number of items.
3. Rotate the factor matrix orthogonally to simple structure.
4. Compute factor scores for each case on each section.

Steps for Principal Components with Analysis of Variance		Steps for Principal Components with Multiple Regression	
ANOVA 5.	Employ analysis of variance on each factor using group memberships as the independent variable and factor scores on the vector as the dependent variable.	Mult R 5.	Employ multiple regression on each factor using group memberships as dummy predictor variables and factor scores as criterion variables.
ANOVA 6.	Identify as possibly biased those items with major loadings on factors which have significant F tests for main effects or for interactions.	Mult R 6.	Identify as possibly biased those items with major loadings on factors which have significant F tests for main effects or for interactions.

The First Method Applied to the Draw-a-Person Test

This study was the third in a series by the present investigator examining the item performance of ethnically different first grade children on the Goodenough-Harris Drawing Test, Man Scale (1963).

Earlier work by Goodenough (1928, 1931, 1950) and by Harris (1963, 1964) indicated that culture systematically affects the performance of children on human figure drawings. In fact, a sex by age by ethnic background interaction would be predicted from these findings if total score on the Draw-a-Man Test (1926) were the dependent variable. Similar expectations can be held for total score performance on the man and the woman scales of the Goodenough-Harris Drawing Test (1963).

Earlier studies by this author (Merz, 1970 and 1971) indicated striking similarities in factor composition of item scores on the man scale of the Goodenough-Harris Drawing Test (DAP), casting doubt on the effects of ethnic differences. Utilizing principle components analysis with orthogonal rotation to simple structure, 12-14 factors for each ethnic group were found; loadings across factors by each group pointed more to similarity among groups than to differences between groups. One could conclude from this that there were similar factor constructs. One could not conclude, however, that the groups did not perform differently.

Method

General Procedure

To investigate whether or not children from diverse cultural backgrounds differ in their item performance on the DAP, another principal components analysis was undertaken. Factor scores were computed for each factor, and analysis of variance was performed on the factor scores. Classification variables for this analysis were ethnic group membership, age, and sex.

Subjects

A total of 1,294 six and seven year old children were used in this study. They belonged to five ethnic groups: Anglo, Black, Mexican-American, Pueblo and Yucca. All were from low-income families attending schools eligible for support under Title I of the Elementary and Secondary Education Act of 1965. These children participated in an oral language development program developed by the Southwestern Cooperative Educational Laboratory in Albuquerque, New Mexico. This testing was part of a larger evaluation effort of the regional laboratory. The children resided in the states of Arizona, New Mexico, Oklahoma, and Texas. A total of 1,402 drawings had been scored for this study;

113 drawings of eight year old, first grade children were not used in the present analysis because these cases created severe disproportionality and empty cells, thus making analysis impossible. From the large number of cases used and its broad geographic representation as well as from general observation and comment, it appeared that the sample was typical of the population of interest. Table 1 describes the composition of the group.

- - - - - - - - - - - - - - - - - -

Insert Table 1 about here.

- - - - - - - - - - - - - - - - - -

Test Employed

The DAP is scored on 73 items; each item performed correctly is given 1 point. DAP's were scored by two laboratory employees under the supervision of the author. First, the scorers were trained to criterion, intertest score reliabilities of .88 to .92. Harris' Children's Drawings as Measures of Intellectual Maturity (1963) was utilized in training. Each scorer studied the man scale procedure by herself and then with the author. Next, sample drawings in Harris' books were scored by each scorer, and any item discrepancies were discussed until mutual agreement on scoring was reached. A second sample of drawings from Harris' book was selected. The same procedure of resolving differences was employed. Finally, several samples of drawings by children included in the laboratory programs during the 1967-68 school year were selected and scored. Reliabilities ranged from .85 on the first Harris drawings to .92 for drawing sample from the 1967-68 school year. Interscorer reliabilities for data included in this study range from a low of .85 for the first 25 drawings to a high of .94 for 50 drawings scored in the last phase of training. These correlations compare favorably with those reported by Harris (1963).

A number of items were deleted because of the statistical analysis employed. Dichotomous variables which have extremely small frequencies of response or extremely large frequencies of response have small variances (Nunnally, 1967). To avoid partitioning error variance items having fewer than 6% of the total sample responding correctly were excluded from the study. Items with 94% of the total sample responding correctly were included in the analysis because they were part of a dependent series; for example, presence of pupils depends upon including eyes. Including eyes was a high frequency item, over 96% correct response. Thirty-four of the 73 items were included in the analysis.

Data Analysis

Principal components analysis was performed with orthogonal rotation to simple solution. Criteria for inclusion in rotation was eigenvalues of 1.0 or more but no more than 15 variables to be included. Factor scores were calculated for each case on each factor. To do this, the BMD X 72 program of the Biomedical Computer Programs, Health Sciences Computing Facility, University of California at Los Angeles. Computation was performed on Stanford University's IBM 360-67 computer.

To avoid the problems associated with disproportionality, a random sample of each group was selected so that there would be 66 pupils for each ethnic group. This undersampled Black and Mexican-American children; yet, it allowed more complete analysis without masking the main effects being studied. Table 2 summarizes the distribution of subjects by ethnic group, age, and sex.

- - - - - - - - - - - - - -

Insert Table 2 here.

- - - - - - - - - - - - - -

Finally, a 5 x 2 x 2 factorial analysis of variance was performed for each of the twelve factors. The following null hypotheses were tested for each factor:

1. There are no significant differences among the means of the factor score for the five ethnic groups.
2. There are no significant differences between the means of the factor scores for the two sexes.
3. There are no significant differences between the means of the factor scores for the two ages.
4. There are no significant first order interactions.
5. There are no significant second order interactions.

The analysis was performed on Stanford University's computer using the general linear hypothesis program BMD X 64.

Findings

Principle Components Analysis

Table 3 presents a summary of the twelve factors extracted; only those variables with loadings of .35 or more are included. The factors appear straightforward; one might label them in this way:

Factor 1 Representation of the Trunk

Factor 2 Eye detail

Factor 3 Facial features

Factor 4 Representation of fingers

Factor 5 Representation of feet

Factor 6 Attachment of head and limbs to trunk

Factor 7 Representation of hair

Factor 8 Representation of limbs

Factor 9 Limb junctures and proportion

Factor 10 Extremities

Factor 11 Proportion

Factor 12 Inclusion of head

- - - - - - - - - - - - - -

Insert Table 3 here.

- - - - - - - - - - - - - -

Analysis of Variance

Table 4 presents a summary of the analysis of variance of factor scores. Relating this information to the five hypotheses; one finds these differences:

Hypothesis 1. Differences among means for ethnic groups

Factor 5. Foot representation $p < .01$

Factor 9. Limb juncture and proportion $p < .05$

Hypothesis 2. Differences between means for the sexes

Factor 9. Limb juncture and proportion $p < .01$

Factor 10. Extremities $p \leq .025$

Hypothesis 3. Differences between means for age

Factor 1. Representation of the trunk $p < .001$

Factor 2. Eye detail $p < .05$

Factor 4. Representation of fingers $p < .05$

Factor 9. Limb junctures and proportion $p < .001$

Hypothesis 4. First order interactions

Factor 1. Representation of the trunk, Ethnic by Age $p < .01$

Factor 5. Representation of feet, Ethnic by Sex $p. < .005$

Factor 11. Proportion, Ethnic by Sex $p < .025$

Hypothesis 5. Second order interactions - none found.

- - - - - - - - - - - - -

Insert Table 4 here.

- - - - - - - - - - - - -

Items loading on each factor identified as possibly biased would be considered biased, as well.

Further Considerations

An analysis using the multiple regression approach was not undertaken with these data; however, reanalysis of the present data with the multiple regression approach is now under way. Two additional analyses are suggested as alternatives. For the principal components analysis and analysis of variance an alternative might be to employ principal components analysis and multiple analysis of variance. For the principal components analysis and multiple regression an alternative might be to enter the dummy categorical variables into the item intercorrelation matrix prior to the principal components analysis; the bias could be estimated directly by the loadings of the dummy variables on each factor. These last methods are being investigated along with the two proposed earlier in this paper.

Reference List

CTB/McGraw-Hill. Technical Bulletin No. 1. Comprehensive Tests of Basic Skills, Form S, 1974.

Echternacht, G. A quick method for determining test bias. Educational and Psychological Measurement, 1974, 34, 271-280.

Goodenough, F. L. Studies in the psychology of children's drawings. Psychological Bulletin, 1928, 25, 272-283.

Goodenough, F. L. Children's drawings. In Murchison (Ed.) A handbook of child psychology. Worcester, Mass.: Clark University Press, 1931.

Goodenough, F. L. and Harris, D. B. Studies in the psychology of children's drawings, II, 1928-1949. Psychological Bulletin, 1950, 47, 369-433.

Green, D. R. Racial and Ethnic Bias in Test Construction. Monterey, CA: CTB/McGraw-Hill, 1971.

Green, D. R. Racial and ethnic bias in achievement tests and what to do about it. Monterey, CA: CTB/McGraw-Hill, 1973.

Green, D. R. What does it mean to say a test is biased? Paper presented at the 1975 Annual Meeting of the American Educational Research Association. Washington, D.C.

Green, D. R. and Draper, J. F. Exploratory studies of bias in achievement tests. Paper at the 1972 Annual Meeting of the American Psychological Association, Honolulu, Hawaii.

Harris, D. B. Children's drawings as measures of intellectual maturity. New York: Harcourt, Brace, World, Incorporated. 1963.

Harris, D. B. Cross-cultural studies of children's drawings. Paper presented at the IX Inter-American Congress on Psychology, 1964. (Dittoed.)

Merz, W. R. Factor analysis as a technique in analyzing test item bias. Paper presented at the 1973 Annual Meeting of the California Educational Research Association, Los Angeles, CA.

Merz, W. R. A biased test may be fair, but then what does bias really mean? Paper presented at the 1974 Annual Meeting of California Educational Research Association, San Francisco, CA.

Ozenne, D. G., Van Gelder, N.C., and Cohen, A.J. Emergency School Aid Act (ESAA) National Evaluation, Achievement Test Restandardization. Santa Monica, CA: Systems Development Corporation, 1974.

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the 1975 Annual Meeting of the American Educational Research Association, Washington, D.C.

Table 1

Distribution of Subjects by Ethnic Origin, Sex, and Age in Total Sample

	Boys		Girls		Total
	Six Years Old	Seven Years Old	Six Years Old	Seven Years Old	
Anglo	31	36	38	29	134
Black	79	119	65	105	368
Mexican-American	158	180	143	174	655
Pueblo	11	24	16	17	68
Yucca	7	26	10	26	69
Totals	286	385	272	351	1,294
	671		623		

Table 2

Distribution of Subjects by Ethnic Origin, Sex, and Age in Subsample

	Boys		Girls		Total
	Six Years Old	Seven Years Old	Six Years Old	Seven Years Old	
Anglo	15	18	19	14	66
Black	13	20	13	20	66
Mexican-American	15	18	15	18	66
Pueblo	10	23	16	17	66
Yucca	7	26	9	24	66
Totals	60	105	72	93	330
	165		165		

Table 3

Summary of Principal Components
Analysis After Orthogonal Rotation

Factor	Item	Loading on Factor	% of Variance Accounted for
1			12
	Arms present	.37	
	Arms and legs attached to trunk	.48	
	Trunk present	.82	
	Trunk represented in two dimensions	.74	
	Limbs represented in two dimensions	.36	
	Any clothing present	.40	
2			6
	Eyes: brow represented	-.38	
	Eyes: pupil represented	-.85	
	Eyes: proportion	-.73	
	Eyes: glance	-.80	
	Nose represented in two dimensions	-.43	
3			6
	Eyes present	-.60	
	Nose present	-.68	
	Mouth present	-.76	
4			4
	Fingers present	-.86	
	Fingers: number	-.50	
	Fingers: detail	-.76	
5			4
	Feet present	.50	
	Feet: proportion	.76	
	Feet: heel represented	.72	
6			4
	Neck present	-.70	
	Limbs: attached to trunk at proper points	-.58	
7			4
	Hair present	-.81	
	Hair on circumference of head	-.72	
	Ears present	-.36	
8			4
	Legs present	-.78	
	Arms and legs attached to trunk	-.60	
	Limbs represented in two dimensions	-.42	

Table 3 (continued)

Factor	Item	Loading on Factor	% of Variance Accounted for
9			3
	Arms at side	-.46	
	Hips: crotch represented	-.71	
	Limbs represented in two dimensions	-.42	
10			3
	Eyes: brow represented	-.50	
	Hands present	-.70	
	Feet present	-.36	
11			4
	Head proportion	-.58	
	Leg proportion	.69	
12			2
	Head present	-.77	
	Arms present	-.44	

Table 4

Summary of Analysis Variance for Each Factor

Factor	Source	df	M S	F	
1	Ethnicity	4	1.272	1.47	
	Sex	1	2.015	2.34	
	Age	1	12.060	13.98	p < .001
	Ethnicity X Sex	4	1.997	2.32	
	Ethnicity X Age	4	3.007	3.48	.01> p > .005
	Sex X Age	1	0.070	0.08	
	Ethnicity X Sex X Age	4	0.838	0.97	
2	Ethnicity	4	1.418	1.40	
	Sex	1	0.010	0.01	
	Age	1	4.056	4.03	.05> p > .025
	Ethnicity X Sex	4	0.766	0.76	
	Ethnicity X Age	4	0.606	0.60	
	Sex X Age	1	0.620	0.62	
	Ethnicity X Sex X Age	4	0.410	0.40	
3	Ethnicity	4	0.462	0.56	
	Sex	1	0.354	0.43	
	Age	1	0.120	0.14	
	Ethnicity X Sex	4	0.600	0.73	
	Ethnicity X Age	4	1.208	1.47	
	Sex X Age	1	0.000	0.00	
	Ethnicity X Sex X Age	4	0.816	0.99	
4	Ethnicity	4	0.466	0.464	
	Sex	1	1.774	1.764	
	Age	1	3.878	3.856	.005> p >.001
	Ethnicity X Sex	4	0.972	0.966	
	Ethnicity X Age	4	0.704	0.670	
	Sex X Age	1	0.024	0.024	
	Ethnicity X Sex X Age	4	2.028	2.016	
5	Ethnicity	4	3.583	3.61	.01> p >.005
	Sex	1	0.475	0.48	
	Age	1	0.150	0.15	
	Ethnicity X Sex	4	3.967	4.00	.005> p >.001
	Ethnicity X Age	4	1.270	1.28	
	Sex X Age	1	0.094	0.10	
	Ethnicity X Sex X Age	4	1.912	1.92	
6	Ethnicity	4	1.399	1.40	
	Sex	1	1.600	1.60	
	Age	1	0.010	0.01	
	Ethnicity X Sex	4	1.048	1.04	
	Ethnicity X Age	4	0.312	0.31	
	Sex X Age	1	1.124	1.12	
	Ethnicity X Sex X Age	4	1.810	1.81	

Table 4 (continued)

Factor	Source	df	M S	F	
7	Ethnicity	4	2.257	2.35	
	Sex	1	0.364	0.38	
	Age	1	1.398	1.46	
	Ethnicity X Sex	4	0.552	0.58	
	Ethnicity X Age	4	0.916	0.95	
	Sex X Age	1	0.256	0.26	
	Ethnicity X Sex X Age	4	0.982	1.02	
8	Ethnicity	4	0.310	0.28	
	Sex	1	0.328	0.30	
	Age	1	0.224	0.20	
	Ethnicity X Sex	4	0.530	0.48	
	Ethnicity X Age	4	1.494	1.36	
	Sex X Age	1	0.791	0.72	
	Ethnicity X Sex X Age	4	1.431	1.30	
9	Ethnicity	4	2.652	2.72	.05> p >.01
	Sex	1	6.818	6.99	.01> p > .005
	Age	1	13.962	14.32	.001> p
	Ethnicity X Sex	4	0.694	0.71	
	Ethnicity X Age	4	2.294	2.35	
	Sex X Age	1	1.242	1.27	
	Ethnicity X Sex X Age	4	1.605	1.64	
10	Ethnicity	4	1.237	1.28	
	Sex	1	4.856	5.04	.025> p >.01
	Age	1	0.341	0.35	
	Ethnicity X Sex	4	0.494	0.51	
	Ethnicity X Age	4	0.097	0.10	
	Sex X Age	1	1.442	1.50	
	Ethnicity X Sex X Age	4	0.125	0.13	
11	Ethnicity	4	0.452	0.394	
	Sex	1	0.010	0.009	
	Age	1	0.340	0.296	
	Ethnicity X Sex	4	3.736	3.256	.025> p > .01
	Ethnicity X Age	4	0.278	0.242	
	Sex X Age	1	0.196	0.170	
	Ethnicity X Sex X Age	4.163	1.163	1.013	
12	Ethnicity	4	1.114	0.708	
	Sex	1	0.196	0.124	
	Age	1	3.624	2.302	
	Ethnicity X Sex	4	1.514	0.962	
	Ethnicity X Age	4	0.950	0.604	
	Sex X Age	1	2.210	1.404	
	Ethnicity X Sex X Age	4	0.534	0.339	